

Mestrado em Humanidades Digitais

Diário de Bordo

Análise e Visualização de Dados

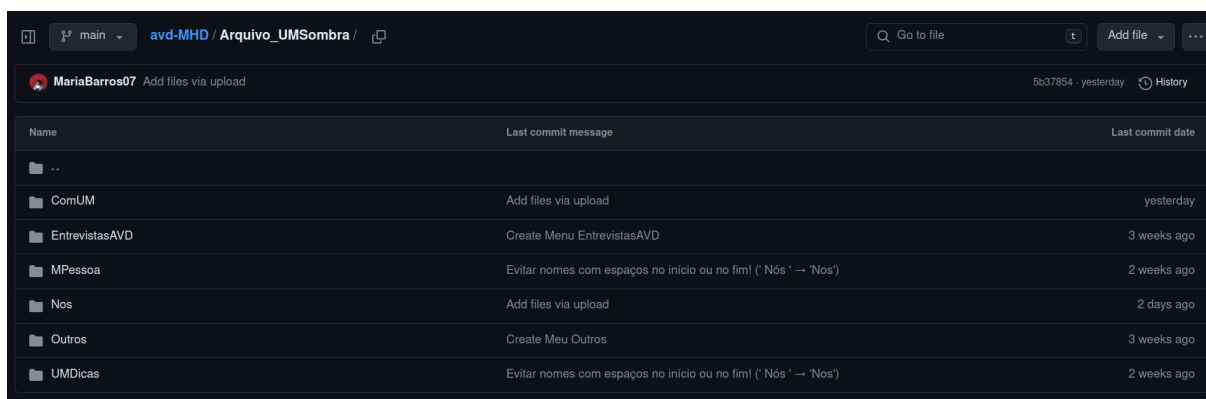
22/03/2024

Gabriela C. Macieira, PG52761

Ano letivo 2023/2024

1 Arrumação do Dataset

Na passada aula de dia 22 de março, foi feita uma visita e análise cirurgica às extrações feitas por cada grupo, de modo a estabelecer normas *standard* a adotar por todos os grupos. Os documentos extraídos e tratados estão armazenados numa pasta comum no GitHub, intitulada de *Arquivo UMSombra*.



Name	Last commit message	Last commit date
..		
ComUM	Add files via upload	yesterday
EntrevistasAVD	Create Menu EntrevistasAVD	3 weeks ago
MPessoa	Evitar nomes com espaços no início ou no fim! ("Nós" → "Nos")	2 weeks ago
Nos	Add files via upload	2 days ago
Outros	Create Meu Outros	3 weeks ago
UMDicas	Evitar nomes com espaços no início ou no fim! ("Nós" → "Nos")	2 weeks ago

Figura 1: Pasta GitHub com as extrações de cada grupo

Tendo previamente visitado o repositório da turma no GitHub, o professor José João deu conta de alguns erros de formatação comuns à maioria dos grupos. Estes erros consistiam essencialmente em duas coisas: *non-breaking spaces* (exemplo assinalado a vermelho na figura 2) e o mau posicionamento dos acentos que, por vezes, aparecem fora da letra à qual devem corresponder. Para além disto, as datas presentes nos metadados de cada grupo, têm formatações diferentes, o que implica uma normalização das mesmas em todos os casos. Consequentemente, o professor ficou encarregue de criar um script que corrija as falhas e que normalize as datas.

```
---
date:30Mai2019
author:Catarina Ferreira
image:https://www.comumonline.com/wp-content/uploads/2019/05/ea0bc75e57985eee9ba9a
1b660f72680w1000.jpg
title: AAUM de prata no nacional universitário de Kickboxing
url: https://www.comumonline.com/2019/05/aaum-de-prata-no-nacional-universitario-d
e-kickboxing/
site: ComUM
description: A Associação Académica da Universidade do Minho (AAUM) ficou no segun
do lugar do pódio atrás da Universidade do Porto (UPorto) e à frente da Associação
tags: AAUM, Kickboxing, Campeonato Nacional Universitário
type: article
---
```

Figura 2: Exemplo de um *non-breaking space*

Utilizando ainda a Figura 2 como exemplo (onde é visível o YAML de um ficheiro Markdown do grupo 2), são visíveis dois pequenos problemas. Em primeiro lugar, existe uma linha em branco antes dos três tracinhos que dão início à secção dos metadados; isto pode causar problemas posteriormente, aquando da conversão do documento para outro tipo de formatação, por exemplo, para PDF.

Em segundo lugar, os dois pontos (:) que sucedem as etiquetas identificadoras dos dados devem ter um espaço antecedente ao texto. Por exemplo, em "date:30Mai2019", os dois pontos que antecedem a data, devem estar espaçados da mesma, devendo ficar "date: 30Mai2019". Este mínimo pormenor faz toda a diferença numa posterior conversão destes metadados.

E foi desta forma que se fez a análise dos ficheiros Markdown conseguidos pela turma. A partir desta análise, foram propostos vários ajustes direcionados às necessidades de cada grupo, pois cada um lida com formatos de dados diferentes. Apesar destas diferenças, existem observações e ajustes que devem ser tidos em conta por todos os grupos sem exceção, que irei especificar de agora em diante.

- Como já mencionei, as datas devem ser normalizadas, de modo a terem um tipo de formatação comum aos grupos.
- Tem lógica que os grupos tenham o mesmo conjunto de metadados mas, em alguns casos, não ajuda nem faz muito sentido. Convém apenas que os conceitos sejam os mesmos em todos, por exemplo, em alguns casos, o entrevistador está definido como *author* e noutros está definido como *interviewer*. Seria benéfico regularizar este tipo de diferenças e talvez será algo a tratar numa fase posterior.
- A secção dos metadados deve estar entre três tracinhos (- - -) como é visível também na Figura 2. Isto faz com que o documento Markdown seja válido. Por esta mesma razão, os tracinhos iniciais não devem ter uma linha em branco a anteceder.

Após a definição destes *standards* e a consolidação de alterações a serem feitas, cada grupo ficou encarregue de ajustar os Markdowns gerados e, a partir daqui, os trabalhos a desenvolver podem prosseguir.